Computationally Efficient Analysis of Large Array FTIR Data In Chemical Reaction Studies Using Distributed Computing Strategy

Ms Suyun Ong, Dr. Wee Chew, Dr. Marc Garland*

Institute of Chemical and Engineering Sciences,
1 Pesek Road, Jurong Island.
SINGAPORE 627833

Keywords: parallel/ distributed computing, chemometrics, Band-target Entropy Minimization (BTEM),
chemical reaction studies, infrared spectroscopy

Abstract

The application of infrared spectroscopy in chemical R&D provides useful chemical group information about chemical transformations. Data acquired from *in situ* infrared measurements of chemical reactions contains spectral profiles that reveal changes in the chemical species composition as reaction time progresses. This spectral information is very useful for exploratory investigations of chemical reactions. Using good experimental design coupled with *in situ* infrared instrumentation, real time data that contains hundreds of thousand, if not millions, of data values are routinely obtained. Such large arrays of data are complex and thus require multivariable statistical analysis to elicit meaningful chemical information. The relatively new MATLAB *Distributed Computing Toolbox* (DCT) and *Distributed Computing Engine* (DCE) was recently employed to enable parallelized distributed computing capabilities for chemometrics analysis of such large *in situ* FTIR data, in particular that from hydroformylation of cyclopentene using rhodium/ rhenium organometallic catalytic precursors. The first parallelized distributed computing version of the Band-Target Entropy Minimization (BTEM) curve resolution algorithm was implemented and applied on aforesaid *in situ* FTIR data from hydroformylation of cyclopentene. This new computing strategy proved to be highly efficacious as more than 10 times reduction in computational time was observed.

Introduction

Groundwork was performed to implement parallel/ distributed computing strategies for solving large array spectroscopic data at the Institute of Chemical and Engineering Sciences (ICES), A*STAR, Singapore [1]. It is presently possible to harness a Window-based computer cluster to implement parallelized distributed computing strategy for chemometrics analysis of large array infrared data. Specifically, the original Band-Target Entropy Minimization (BTEM) curve resolution technique [2, 3]

was re-adapted into its parallelized distributed computing analogue using the MATLAB® *Distributed Computing Toolbox* (DCT) and *Distributed Computing Engine* (DCE).

The MATLAB DCT and DCE are relatively new products released by MathWorks, Inc. [4]. They allow the user (or client) to program parallel/ distributed computing MATLAB applications and execute them in computer clusters without having to leave the deployment area. Through the MATLAB DCT/ DCE environment, the user PC (client) uses DCT commands to create jobs to be sent to the server's job manager (i.e. cluster head node), which will then assign tasks to its various workers via DCE. Once the calculated results are generated by every worker, they will be collated and sent back to the client via the job manager.

The efficacy of parallelized distributed computing via MATLAB DCT/ DCE platform was tested on a large array of *in situ* FTIR spectroscopic data obtained from hydroformylation of cyclopentene using rhodium/ rhenium organometallic catalytic precursors. This dataset was previously analyzed in the IBM-IHPC High Performance Computing (HPC) Quest competition in 2003 [5]. The *serial processing* strategy (i.e. one parametric BTEM computation after another) was used in the IBM-IHPC Quest competition, with all computations performed on a supercomputing cluster located at the Institute of High Performance Computing (IHPC), A*STAR, Singapore. This paper presents a comparative study between (i) the new *parallelized* distributed computing BTEM implementation on a Window-based cluster at ICES, (ii) a typical BTEM *serial processing* on a single IBM personal computer, (iii) and the *serial processing* results from 2003 IBM-IHPC High Performance Computing (HPC) Quest competition.

Results and Discussion

In the present work, all parallelized distributed computing jobs (i.e. individual parametric BTEM runs) are deployed using a in-house written Graphical User Interface (GUI), ***runDCT* GUI** (see Figure 1), which runs on the MATLAB DCT client personal computer. The jobs are then sent to the cluster server job manager via MATLAB DCE (see Figure 2).

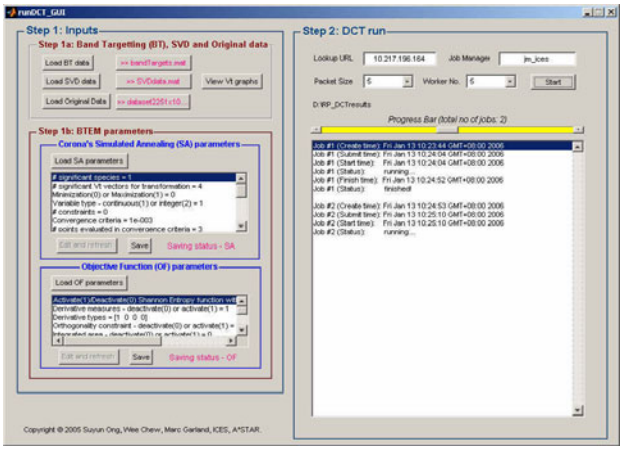


Figure 1 runDCT GUI – Running BTEM on MATLAB DCT/ DCE Distributed Computing Platform

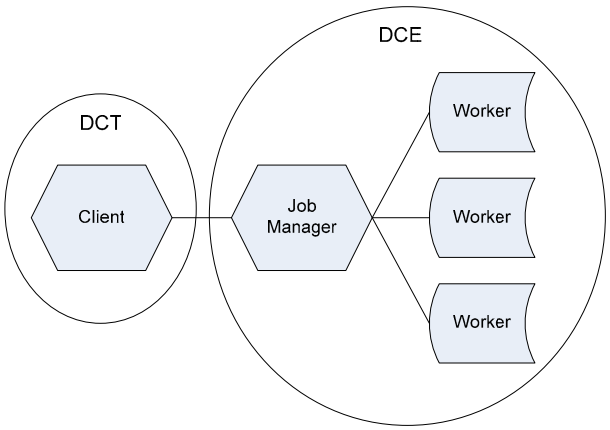


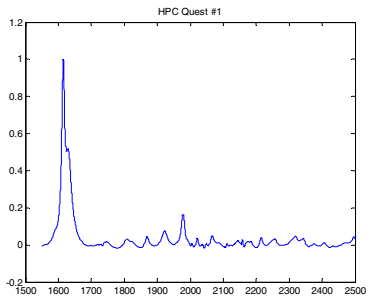
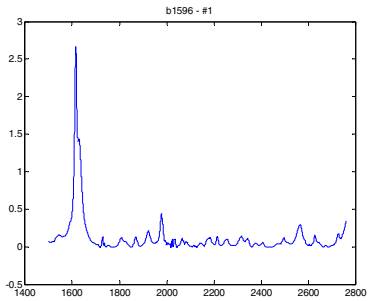
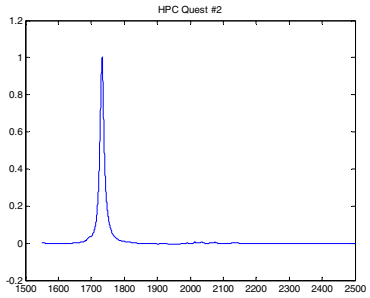
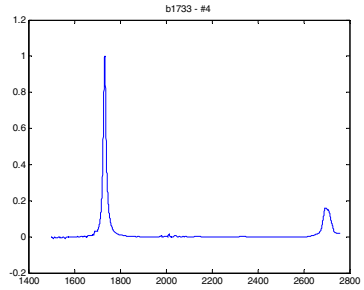
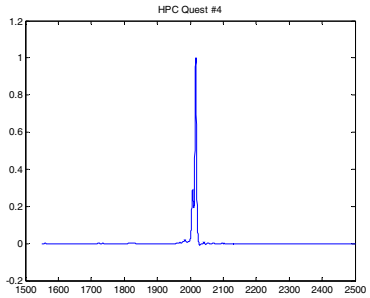
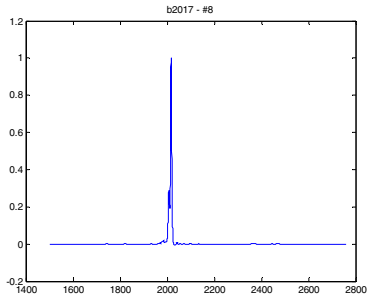
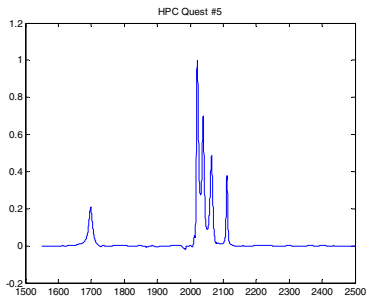
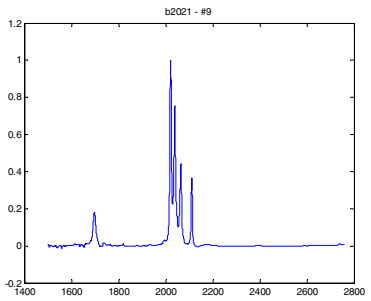
Figure 2 Basic Distributed Computing Configuration using MATLAB DCT/ DCE

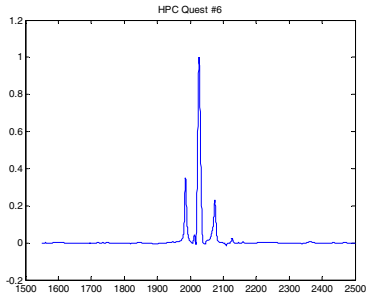
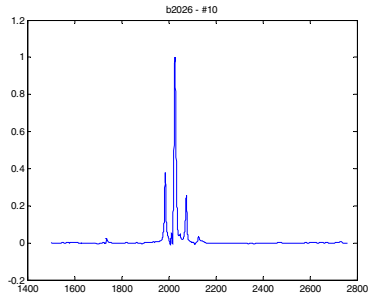
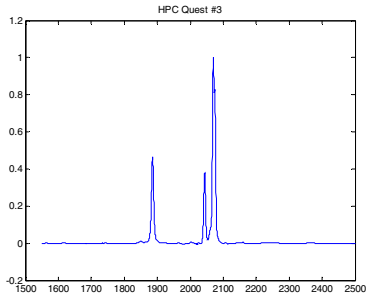
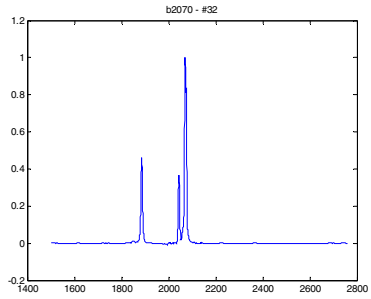
From the series of rhodium/ rhenium catalyzed cyclopentene hydroformylation experiments, a total of 2548 *in situ* FTIR spectra with a 6301 channels of wavenumbers for the range 2760-1500 cm^{-1} was analyzed. All spectra were collated into a single data matrix $\mathbf{A}_{2548\times6301}$, and its singular value decomposition was subsequently performed. From the abstract right singular V^T vectors of $\mathbf{A}_{2548\times6301}$, 50 band-targets were visually identified. Each of these 50 band-targets will individually undergo BTEM pure component spectra reconstruction via (i) *serial processing* on a single IBM X36 PC with 512 MB RAM and (ii) *parallelized* distributed computing through *runDCT* GUI, with a IBM XSERIES 346 having Intel® Xeon™ CPU of 3.20 GHz and 4 GB RAM as *DCT head node* and DCE server with *12 worker nodes*, each having a IBM Intel Pentium® 4 CPU of 3.40GHz and 1 GB RAM. The Corana's Simulated Annealing (SA) and Objective Function (OF) parameters in the *runDCT* GUI BTEM computations were varied as delineated below, with one set of parameters similar as those used in the 2003 IBM-HPC Quest competition for one-to-one comparison.

- Varying significant $\mathbf{V}^T$ vectors for transformation e.g. z = 5, 10, 25, 30, 40, 50 and 75
- Varying number of DCE workers e.g. w = 4, 8 and 12, and packet size p sent via DCT ($w = p$)
- Activate second spectral derivative minimization term in BTEM objective function
- Deactivate non-negative concentration constraint

The pure component spectra reconstructed in the parallel *runDCT* BTEM computations are compared with those of the 2003 IBM-HPC Quest competition (see Table 1). The comparison of computational time is provided in Table 2 below.

Table 1 Comparison of Resolved Pure Component BTEM Spectral Estimates

Band	Compound	HPC Quest Serial BTEM	Parallel *runDCT* BTEM
Band #1 b1596	cyclopentene	#1 	$z = 30$ 
		#2	$z = 10$
Band #4 b1733	cyclopentane carboxaldehyde		
		#4	$z = 30$
Band #8 b2017	$HRe(CO)_5$		
		#5	$z = 10$
Band #9 b2021	$RCORh(CO)_4$		

Band	Compound	HPC Quest Serial BTEM	Parallel *runDCT* BTEM
		#6	$z = 10$
Band #10 b2026	$RhRe(CO)_9$		
		#3	$z = 25$
Band #32 b2070	$Rh_4(CO)_{12}$		

z	BTEM Computational Time (hh:mm:ss)				
	IBM X36 PC	2003 IBM-IHPC HPC Quest	*runDCT* GUI		
			*w*4*p*4	*w*8*p*8	*w*12*p*12
5	1:22:36	-	0:14:42	0:09:30	0:08:57
10	2:24:21	-	0:33:13	0:18:36	0:14:02
25	6:11:55	8:44:24 †	1:16:10	0:49:29	0:35:09
30	7:58:20	-	1:53:42	1:02:22	0:44:55
40	10:43:17	-	2:26:15	1:28:58	1:04:42
50	14:36:59	-	3:42:03	2:01:48	**
75	25:47:38	-	**	**	**

Table 2 Total Computational Time required for BTEM runs in different hardware/ software platforms

† In the HPC Quest, the CPU time of 8:44:24 included the time to automatically find 60 band-targets and their subsequent *serial processing* BTEM computations. The computational time for *runDCT* runs includes only BTEM computations on 50 visually identified band-targets. The reduction in computational time has to take this difference into account.

** Indefinite computational time was observed and hence the runs were aborted.

From this comparative study, the efficacy of the first parallelized BTEM computational strategy

implemented via MATLAB DCT/ DCE platform is well demonstrated. The *runDCT* BTEM pure component spectral estimates were strikingly similar to those obtained previously in the 2003 IBM-IHPC HPC Quest competition, and its computational time for $z = 25$ V^T vectors using 12 DCE workers displayed more than 10 times decrease in total computational time. Though several *runDCT* trial executions with $z = 50$ and 75 V^T vectors were aborted because of indefinite computational time, it is still hopeful that more can be achieved through parallelization of computation algorithms via MATLAB DCT/ DCE for large array spectroscopic data analysis in chemical reaction studies.

References

[1] S. Y. Ong, *Implementation of Distributed Computing Strategy for Chemometrics Analysis of Multidimensional Spectroscopic Data*. Final Year Research Project Thesis. National University of Singapore, 2006.

[2] **Error! Reference source not found.**

[3] **Error! Reference source not found.**

[4] Distributed Computing Toolbox, Users Guide, Version 1. © 2004-2005 by The MathWorks, Inc.

[5] M. Garland, Chen, L., Li, C. Z., Zhang, H. J., Chew, W. and Widjaja, E. (2003). *Massively Parallel Entropy Based Pattern Recognition for System Identification in Noble Metal Mediated Chemical Syntheses (Project: HPC 035)*, Final report for the 2003 IBM-IHPC High Performance Computing Quest competition.